

DOCUMENT RESUME

ED 177 190

TM 009 633

AUTHOR McCullagh, Peter
TITLE Analysis of Ordered Categorized Data. Technical Report No. 62. Revised.
INSTITUTION Chicago Univ., Ill. Dept. of Statistics.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Jun 78
GRANT NSF-SOC76-80389
NOTE 64p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Data Analysis; Goodness of Fit; *Mathematical Models; *Multiple Regression Analysis; Scores; *Statistical Analysis; *Statistical Data
IDENTIFIERS *Categorical Data; Log Linear Models; Qualitative Data

ABSTRACT

A general class of regression models for ordered categorised data is developed and discussed. Their properties are contrasted with log linear models and other methods of analysis. The general problem of fitting by maximum likelihood is discussed and solved. The models are shown to be multivariate extensions of generalised linear models and the fitting method is iteratively re-weighted least squares. Special attention is given to a particular model--the cumulative logit model--and alternative fitting methods are described for this case. Applications are discussed with the aid of three examples. (Author/GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ANALYSIS OF ORDERED CATEGORISED DATA*

by

Peter McCullagh

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Peter McCullagh

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

TECHNICAL REPORT NO. 62

Department of Statistics

University of Chicago

Chicago, Illinois 60637

June 1978

Revised: October 1978

* This research was supported in part by National Science Foundation Grant No. NSF SOC76-80389.

ABSTRACT

A general class of regression models for ordered categorised data is developed and discussed. Their properties are contrasted with log linear models and other methods of analysis. The general problem of fitting by maximum likelihood is discussed and solved. The models are shown to be multivariate extensions of generalised linear models and the fitting method is iteratively re-weighted least squares. Special attention is given to a particular model — the cumulative logit model — and alternative fitting methods are described for this case. Applications are discussed with the aid of three examples.

Some key words: generalised empirical logit transform; jackknife; link function; location parameter; log-linear model; logit linear model; multivariate generalised linear model; ordered categories; scale parameter; scores.

Acknowledgements

Much of the work for this paper was carried out for the author's Ph.D. thesis at Imperial College, London, where financial support was provided by the Health and Safety Executive. I wish to thank

F. J. Anscombe,

A. C. Atkinson,

D. R. Cox,

L. A. Goodman,

S. J. Haberman,

W. H. Kruskal,

R. L. Plackett,

and D. L. Wallace

for helpful discussions and constructive comments on earlier versions of this paper.

Introduction

It is widely recognised that the types of data, as well as the class of problems, which a statistician is likely to encounter vary greatly with the field of research. Consequently, methods which are useful in one field of research may be of little use or interest to researchers in another field. In the physical sciences for example, the overwhelming proportion of data are essentially quantitative measurements although the scale on which the measurements are taken may be rather arbitrary. In the social sciences, and to a lesser extent in the biological sciences, qualitative data are more common. It is useful, therefore, to distinguish a few different types of data.

- (a) continuous
- (b) essentially continuous but exhibiting some granularity
- (c) discrete but having a large number of possible values
- (d) censored or grouped into categories with known boundaries
- (e) data fall into a few broad qualitative categories
- (f) only two possible values for each observation

Attributes of individuals are conveniently described by observations of types (e) and (f) and therefore these types of data are usually called qualitative. The present paper is concerned solely with data of type (e) and by implication (f) since this is a special but important example of (e).

The distinguishing feature of the other types of data is that they are quantitative. It is important to distinguish between quantitative categorised data of type (d) and qualitative categorised data (e). Methods for dealing with quantitative categorised data have been developed by Haitovsky (1974) and, in the context of survival data, by Cox (1972).

For clarity, we subdivide qualitative data into several classes and identify the types which can usefully be handled by methods currently available. It is not suggested that either of the subdivisions given here are mutually exclusive or exhaustive. Indeed, survival data is typically a mixture of (a) and (d). The following subdivision for qualitative data is helpful in choosing an appropriate model.

- e(i) nominal - no order among categories
- e(ii) ordinal with known associated scores or quantitative factor
- e(iii) ordinal with underlying continuum but no known associated scores
- e(iv) ordinal but no underlying continuum or associated scores
- e(v) complex order structure

A related classification of scale types and their relevance to the statistical procedures employed is discussed by Stevens (1951, 1958, 1968) who distinguishes nominal, ordinal, interval and ratio scales. An interval or, more correctly, an equal-interval scale can be treated as a special case of e(ii) where the associated scores are the positive integers. Similarly, a ratio scale can be converted to an interval scale by a log transform or it can be considered as a special case of e(ii) where the scores are increasing powers of 2, say. This paper is concerned principally with data of type

(iii) — i.e. ordinal categories with an underlying unobserved or unobservable continuum but no known score or quantitative factor associated with each category. However, some of the other types are considered briefly in section 2.

Data of type (ii) are similar to those described in (d) above. However they might not be truly quantitative since, for example, the scores might be known to lie in arithmetic progression with unknown step length and unknown initial value. It is usually sufficient in (ii) above that the scores be known up to a location-scale transformation. Alternatively, it might be known that the categories form an equal interval scale on some continuum.

In general, it is extremely difficult to distinguish between data of types (iii) and (iv) above since any distinction hinges on the existence or otherwise of an underlying unobserved continuum. No distinction is offered here other than to point out that the models described here can have an interpretation in the absence of any continuum.

Complex order structures might arise, for example, if the observed categories form a cross-classification with row and column categories of types (ii) - (iv). Thus the underlying continuum, if one exists, would be two dimensional. Data of this type do arise in practice but there seems to have been little attempt at developing models for them. One possibility might be to develop models based on the so-called cross ratio distributions, Pearson (1913), Plackett (1965).

2.0 Models for Ordered Categories

2.1 Historical Note.

Many, probably a substantial majority of categorisations have an implicit order. It is ironic therefore, that the spate of recent work on categorised data — mainly developments of log-linear models — should have been concentrated almost exclusively on the nominal case. There are a few notable exceptions mentioned later, but it is important to examine the reasons for the present imbalance. For an excellent short review see Fienberg (1975).

The present imbalance has its roots in a long and bitter dispute between Karl Pearson and G. Undy Yule during the first two decades of this century. Pearson's approach was to view all cross-classifications as arising from a partition of a set of multivariate data and insisted that it was always meaningful to view the data as such regardless of the nature of the categories involved. This extreme viewpoint led him, by analogy with the bivariate normal distribution, to develop the so-called tetrachoric correlation coefficient and its polychoric generalisation as a measure of the association or correlation between two dichotomous or polytomous variables. As a measure of association, the tetrachoric correlation coefficient suffered from two major disadvantages. From the practical viewpoint its evaluation posed problems beyond the computational ability of many potential users. From the theoretical viewpoint, the defects, as pointed out by Yule (1912), hinged mainly on the fact that altering the marginal frequencies in a 2×2 table, while keeping the conditional probabilities in the body of the table constant, changes the tetrachoric coefficient (and also the product moment correlation coefficient) but not measures of association based on the cross product ratio.

Yule's philosophy, on the other hand, was to regard the categories as fixed and to examine the data for structural relationships among the cell frequencies. This approach makes no direct use of order among the categories even when an order exists. The same philosophy led to the work of Bartlett (1935) and forms the basis of the modern Poisson-log-linear model approach. For the 2×2 table Yule's approach led to measures of association based on the cross product ratio which suffered from neither of the disadvantages of Pearson's tetrachoric coefficient. To contrast his and Pearson's philosophy, Yule used an example concerning the association between vaccination and small-pox asserting "All those who are dead are equally dead: no one of them is more dead or less dead than another and the dead are quite distinct from the survivors". No doubt he was correct, but even here there are degrees of survivorship. In any case, the examination of structural relationships among the cell probabilities led to the development of log-linear models and Pearson's ideas were largely ignored.

It seems to me that the neglect of Pearson's views has been a mistake since categories often do represent partitions of an underlying continuum. Recently, there has been a trend towards incorporating order information via Pearson's approach in the work of Snell (1963), Williams & Grizzle (1972), Clayton (1974) and Simon (1974).

It should, perhaps, be emphasized that the view adopted in the present paper is rather less extreme than Pearson's. We do not contend that all polytomies can meaningfully be regarded as contiguous intervals on a continuum. Instead, if there is an order, consideration of an underlying continuum leads to structural relationships among the cell probabilities different from the more usual log-linear models. The parameters of such a structural model have a simple interpretation in terms of the underlying continuum if one exists. If not, then the model defines a simple structural relationship among the cell probabilities which is often more parsimonious than can be achieved by use of log-linear models. Thus the present approach, therefore, is in keeping with both Yule's and Pearson's views.

2.2 Need for incorporating order into models

The need for models which make use of category order has long been recognised. Yates (1948) points out that the usual chi-squared test for independence covers all forms of departure from proportionality. If the nature of the data is such that departures of a particular type are to be expected, then a test of significance appropriate to departures of this type is justified. Yates' solution, in the absence of information to the contrary, was to assign equally spaced scores to the categories implicitly assuming an interval scale. This method gives valid and useful tests but from the modelling viewpoint it suffers from the arbitrary assignment of equally spaced scores.

Tukey (1954), in his paper on unsolved problems of experimental statistics asks "How soon will we appreciate that the columns (or rows) of a contingency table usually have an order? When there is an order, shouldn't we take this into account in our analyses? How can they be efficient otherwise?" To the latter question we might reasonably add "how can they (the models) be meaningful otherwise?"

In the same issue of J.A.S.A., Goodman & Kruskal (1954) discuss measures of association for cross-classifications. In addition to emphasising the different interpretations of association they point out that a measure of association should be chosen with due regard for order, if any, in the polytomies of the classification. However, they were interested in measures of association rather than in models per se.

On the other hand, Savage's (1957) assertion that "I know of no reason to limit statistical procedures to those involving arithmetic operations consistent with the scale properties of the observed quantities." should not be confused with the genuine need for models reflecting the scale properties of the observed quantities. Thus even if the model is consistent with the scale properties of the observations, the fitting and testing of that model may involve arithmetic operations apparently inconsistent with the scale properties of the observations. For example, in comparing two groups of students with respect to their IQ's it may be necessary to add the observations, but two individuals each with an IQ of 60 can hardly be considered equivalent to one with an IQ of 120. Savage's comment, therefore, does not conflict with the aims of this paper.

More recently, log linear models have been extended so that account can be taken of scores associated with the categories. See for example Haberman (1974), Williams & Grizzle (1972) and Nelder & Wedderburn (1972).

2.3 Some Examples

To illustrate the types of data we wish to consider two examples are given. The first is a table from Maxwell (1961, p. 70) also analysed by Nelder & Wedderburn (1972), giving the number of boys with four different ratings for disturbed dreams in five different age groups. The ratings are on

Table 1

Frequency of disturbed dreams among boys aged 5 - 15.

Age (yrs.)	<u>Rating</u>				Total
	4	3	2	1	
5 - 7	7	3	4	7	21
8 - 9	13	11	15	10	49
10 - 11	7	11	9	23	50
12 - 13	10	12	9	28	59
14 - 15	3	4	5	32	44
Total	40	41	42	100	223

an ordered scale, the higher the rating the more the boy suffers from disturbed dreams. The rows of the table, on the other hand are indexed by a quantitative factor namely the age range of boys in that group.

Table 2 taken from Bradley, Katti & Coons (1962) gives the response frequency of judges in a taste testing experiment. The five possible responses are on an ordered scale ranging from terrible to excellent, while the rows are indexed by an unordered qualitative factor with five levels representing the five treatments.

Table 2

Response frequency in a taste-testing experiment

Treatment	<u>Response Category</u>					Total
	1	2	3	4	5	
1	9	5	9	13	4	40
2	7	3	10	20	4	44
3	14	13	6	7	0	40
4	11	15	3	5	8	42
5	0	2	10	30	2	44
Total	41	38	38	75	18	210

Our approach towards analysing these data is to examine the effect of the other classification or classifications on the response variable. Like regression analysis this is an asymmetric viewpoint and is in sharp contrast to the log-linear model approach — see §2.7. What, for example, is the predicted or estimated proportion of boys aged 9-10 who have severely disturbed dreams? Note that the age range 9-10 does not correspond to a category but the question is, nevertheless, a sensible one. Since we are interested only in proportions of the row totals our inference is conditional on these totals but not on the column totals. Thus the row totals can be fixed as in table 2 or random as in table 1. The column totals, on the other hand, must be random.

For table 2 we might be interested in one of the following:

- (i) identifying the treatment with the most favourable mean response;
- (ii) identifying the treatment for which the judges were most/least in agreement;
- (iii) determining whether the quality of treatment, as measured by the mean response, is related to agreement between judges.

These data are analysed in section 4.

2.4 Fisher's Scoring Method.

In his book, *Statistical Methods for Research Workers*, Fisher (1958, p. 290) described a method for analysing qualitative data. Twelve samples of human blood tested with twelve different sera gave reactions represented in a 12×12 table by five ordered symbols. His method consisted of choosing numerical values for the five symbols in such a way that the ratio of the error sum of squares to the total was minimised. The data were then analysed as if they were continuous, but appropriate allowances were made to the degrees of freedom in the analysis of variance table - three being added to the treatments (rows + columns) and the residual degrees of freedom being reduced by three. An analysis of covariance was performed to determine whether the empirically estimated scores differed significantly from arbitrarily chosen but equally spaced scores. They did not.

One of the major problems associated with Fisher's scoring method is that the scores obtained are not necessarily monotone although if the systematic part of the data is dominant over the random part, the scores tend to be monotone or nearly so. The problem of obtaining the best monotone scores has received considerable attention in the psychological literature. See, for example, Bradley et al. (1962) and Nishisato & Arri (1975).

A second problem is that, with so few different values possible, the data are not nearly normal and this raises doubts about the adequacy of approximations based on the normal and related distributions. Finally there is the problem that the scores are not like parameters in the usual linear model. They are not linear functions of the data and when they are estimated they are treated as if they were the data which they are not.

Despite these conceptual and distributional difficulties the scoring method is often useful as an exploratory technique particularly when more complex order structures exist among the categories. See, for example, Healy & Goldstein (1976).

2.5 Scores in Log-linear Models

The purpose of log linear models is to identify structural relationships such as independence among the cell means in a contingency table. It is found that many of the useful structural models are such that the log of the cell probabilities can be expressed as a linear combination of certain parameters. If the rows and columns of a contingency table are indexed by qualitative factors the only sensible, a priori log linear models are (i) main effects only — i.e. independence — and the rather trivial model (ii) main effects + interaction where all the observations are fitted exactly. In practice, more parsimonious models, intermediate between (i) and (ii), which describe a variety of types of interaction would be more useful. The situation is analogous to the randomised blocks design with several observations per cell where, if interaction exists, we might seek a more parsimonious description of the data by using a suitable transformation or link to give an additive model with no interaction.

If, however, the rows and columns are indexed by a quantitative factor with known values such as the rows of table 1, or a factorial arrangement of several qualitative or quantitative main effects, then a variety of models, intermediate between independence and complete interaction, can be formulated. Thus if known values or scores are attached to the row and column categories so that they are of type (ii) the interaction can often be explained simply by a linear $\times$ linear or linear $\times$ quadratic term (Haberman, 1974). The structural log-linear model for linear $\times$ linear interaction is

$$\log(\pi_{ij}) = \mu + \alpha_i + \beta_j + \gamma(r_i c_j) \quad (2.1)$$

where r_i and c_j are the known scores associated with row i and column j . A similar model was used by Mandel (1959) to explain certain types of interaction in a Latin Square design.

In practice, however, scores are rarely given but this is no real drawback since if a linear $\times$ linear interaction term exists analogous to that in (2.1) a suitable model might be

$$\log(\pi_{ij}) = \mu + \alpha_i + \beta_j + \gamma_i \delta_j \quad (2.2)$$

where either or both of the γ and δ terms are unknown. Obviously, if the categories were ordered we would expect the γ and δ terms to indicate this, but, in principle, a model such as (2.2) could be used when a more complex but qualitative order structure exists among the categories. When both γ and δ are unknown, the model is non-linear and is equivalent to one of the models discussed by Goodman (1978). In a rather different context, Mandel (1971) uses similar mixed additive-multiplicative model to examine the nature of interactions in randomised blocks and related designs.

The usual practice when the scores are not given but we wish to utilise the underlying order, is to use integer scores as if they were given. This is equivalent to setting $r_i = i$ and $c_j = j$ in (2.1) although quadratic and higher terms might be required to obtain an adequate description for the data. The practice of using equally spaced integer scores is usually quite successful since, if the true model is (2.2) with γ and δ both monotone, and we approximate (2.2) by (2.1) with $r_i = i$ and $c_j = j$ the adequacy of the approximation depends on how close γ and δ are to r and c respectively. The space of possible monotone vectors δ is a cone in k -dimensional space and c is the vector at the 'centre' of the cone. The correlation between c and an arbitrarily chosen δ is at least $\sqrt{3/(k+1)}$. Hence if the number of categories is 5, say, the correlation between the true parameter δ and the substituted integer scores is at least 0.71. This analysis suggests that if the number of

categories is small it will usually be difficult to discriminate between an estimated $\hat{\delta}$ and equally spaced scores $\underline{c}$. However, with extensive data it should be possible to find examples for which the equally spaced scores are not adequate to explain the interaction by a linear $\times$ linear term. In such cases, (2.2) could be a useful alternative to using higher order terms.

Abelson and Tukey (1963) have considered the general problem of finding a contrast vector $\underline{c}$ which maximises the minimum correlation with vectors $\hat{\delta}$ subject to $\delta_1 \leq \delta_2 \leq \dots \leq \delta_k$. Their analysis shows that the optimum $\underline{c}$ is widely spaced in the tails and more closely grouped in the centre. This fact suggests that percentile scores based on some symmetric unimodal reference distribution might be a good approximation to the optimum $\underline{c}$, see Williams and Grizzle (1972). However these scores have the disadvantage of being estimated from the data and used as regressors in the model as if they were fixed. The consequences of this assumption do not seem to have been investigated.

Finally we emphasise an important invariance property of log-linear and log-non-linear models such as (2.1) and (2.2). The property is permutation invariance which means that if the rows or columns are permuted then models (2.1) or (2.2) still apply with the appropriate permutations applied to the elements of $\underline{\alpha}$, $\underline{\beta}$, $\underline{\gamma}$ and $\underline{\delta}$. This is a general property of log-linear models and more generally, of models which do not use an order-dependent transformation or link function. A class of order dependent models is described in section 2.6.

2.6 Using an implicit continuum.

We assume that one of the variates — the row variate in examples 1 and 2 — represents contiguous intervals on some implied continuous scale. Thus in example 1 we are interested in a dependent implied variable representing degree or severity of suffering from disturbed dreams. In the second example, the implied dependent variable is the judges opinion for a given treatment.

Of course, if the categories can be represented as contiguous intervals on some scale they can, by a monotone transformation, be represented as contiguous categories on any scale. Incidentally, treating the categories as contiguous intervals, implies that observations in the same category are not identical — they are merely not distinguished due to the crudeness of the scale. We choose a scale such that, for fixed values of the explanatory variables, the continuous dependent variables, if they could be observed, have fixed variance and a mean which is a linear function of the dependent variables. The choice of distributional shape is somewhat arbitrary but the logistic distribution has been chosen from considerations of symmetry and convenience and because logits are preferred by many authors to probits for the analysis of binary data. Since the choice of variance or scale parameter is arbitrary we set the scale parameter of the logistic distribution equal to unity.

Thus if the number of categories or intervals is k and the boundaries between the intervals are $\theta_1, \dots, \theta_{k-1}$, the probability γ_{ij} of an observation Y being category j or less given explanatory variables X_i is

$$\gamma_{ij} = \text{pr}\{Y \leq j \mid X_i\} = \exp(\theta_j - \beta^T X_i) / [1 + \exp(\theta_j - \beta^T X_i)],$$

where β is a vector of unknown parameters.

It is assumed that the observations within a given row of the contingency table are independent and identically distributed so that, conditional on the row sum, they are a multinomial sample with probabilities π_{ij} , $j = 1, \dots, k$, $i = 1, \dots, t$ and cumulative probabilities

$$\gamma_{ij} = \pi_{i1} + \dots + \pi_{ij} \quad (2.3)$$

$$\text{and } \gamma_{ik} = 1.$$

Hence, consideration of the underlying continuum yields the linear structural model

$$\log\{\gamma_{ij}/(1 - \gamma_{ij})\} = \theta_j - \beta^T X_{i1} \quad \begin{matrix} j = 1, \dots, k-1 \\ i = 1, \dots, t \end{matrix} \quad (2.4)$$

In (2.4), $\beta = 0$ implies no association and is equivalent to a log linear model with main effects only. The negative sign in (2.4) ensures that large positive values of $\beta^T X_{i1}$ imply that the higher numbered categories have relatively greater probability and large negative values indicate that the lower numbered categories are relatively more probable. This convention is used to simplify interpretation. It is often convenient to relax the assumption of constant variance yielding the model

$$\log\{\gamma_{ij}/(1 - \gamma_{ij})\} = \{\theta_j - \beta^T X_{i1}\}/\tau_i. \quad (2.5)$$

Small values of τ_i indicate that the observations are tightly grouped; large values indicate that they are widely dispersed. For numerical purposes it is convenient to impose a structure on the scale parameters. This is done in §5.1.

Log linear models for multi-factor contingency tables are symmetric in that no one factor takes precedence over any other. No factor is considered dependent and the others explanatory: all are treated on an equal footing. Logit linear models for binary data are asymmetric: one binary factor is selected for special treatment and the others are explanatory. The analogous procedures for continuous data are respectively correlation analysis which is symmetric and regression analysis which is not. Oddly enough, the logit linear model for binary data occurs as a special case of a log-linear model. It is tempting, therefore, to discard the logit linear model for binary data as merely a special case of the more general log linear model. This would be a mistake as the philosophies behind the two approaches are very different. Furthermore, the natural extension of logit linear models described in this paper is also asymmetric and does not, except in the special case $k = 2$, correspond to a log-linear model.

Choosing which variable to regard as dependent is usually easy as in Tables 1 and 2. Table 3 in §3.4 is less clear but the most natural causal relation is one way: tonsil size may be affected by the presence or absence of *Streptococcus pyogenes*. Consequently we regard tonsil size as dependent and carrier vs non-carrier as an explanatory factor with two levels. For predicting carrier vs non-carrier given tonsil size, we could analyse the data the other way around but only if the sample proportions estimate the population proportions of carriers in the different tonsil size categories — i.e. sampling from a bivariate population. From an epidemiological viewpoint this would mean that the carrier rate in the population for which prediction is required must be the same as the carrier rate in the population sampled — often a most unreasonable assumption. More generally, the choice of a dependent variable may depend on possible causal or chronological relationships. Often it also depends on how the data are going to be used and on the sampling scheme by which they were obtained.

3.1 The Two Sample Problem.

The two sample problem where counts are summarised in a $2 \times k$ contingency table with two ordered columns arises frequently in practice. Although the model for this problem could be regarded as a special case of the general logit linear regression model (2.4) there is sufficient simplification to merit separate treatment.

The linear logistic model for two samples is

$$\begin{aligned}\lambda_{1j} &= \log\{\gamma_{1j}/(1 - \gamma_{1j})\} = \theta_j - \frac{1}{2}\Delta \\ \lambda_{2j} &= \log\{\gamma_{2j}/(1 - \gamma_{2j})\} = \theta_j + \frac{1}{2}\Delta, \quad j = 1, \dots, k-1,\end{aligned}\tag{3.1}$$

where λ_{ij} is called the j^{th} logit for the i^{th} group and Δ is the common difference between corresponding logits. In the following sections we consider the problem of estimating Δ and its standard error. Significance tests concerning Δ can be carried out via the likelihood ratio chi-squared statistic or by comparing an estimate of Δ with its standard error. The latter approach is often preferable especially if one sided tests are required. In addition, the parameter Δ gives a measure of the size of the logistic difference between the two groups as opposed to merely the statistical significance of that difference.

3.2 An Odds-Ratio Estimator of Δ .

Let the cell counts be $\{n_{ij}\}$ with row totals $n_{i\cdot}$, $n_{\cdot j}$ and column totals $\{n_{\cdot j}\}$. The cumulative row sums are denoted by $\{R_{ij}\}$ so that $n_i = R_{ik}$. Since we have assumed multinomial sampling in each row, the marginal distribution of R_{ij} conditional only on the row total n_i is binomial with index n_i and parameter γ_{ij} . Hence the j^{th} sample logit

$$\bar{\lambda}_{ij} = \log\{(R_{ij} + \frac{1}{2}) / (n_i - R_{ij} + \frac{1}{2})\} \quad (3.2)$$

has expectation $\lambda_{ij} + O(n_i^{-2})$ (Cox (1970) p. 33, Plackett (1974) p. 3). Thus, for any fixed weights $\{w_j\}$ with $\sum w_j = 1$, the weighted combination $\tilde{\Delta} = \sum w_j (\bar{\lambda}_{2j} - \bar{\lambda}_{1j})$ has mean $\Delta + O(n_1^{-2} + n_2^{-2})$. For a similar estimator of the parameter Δ , Clayton (1974) derived a formula for the weights which minimise the asymptotic variance of $\tilde{\Delta}$ when $\Delta = 0$. The weights are given by

$$w_j \propto \gamma_j (1 - \gamma_j) (\pi_j + \pi_{j+1}) \quad (3.3)$$

where γ_j is the common value of γ_{1j} and γ_{2j} under the hypothesis that $\Delta = 0$. Using these weights, the asymptotic variance was shown to be

$$\text{Var}(\tilde{\Delta}) = \left[\frac{n_1 n_2}{n} \sum_{j=1}^{k-1} \gamma_j (1 - \gamma_j) (\pi_j + \pi_{j+1}) \right]^{-1} + O(\Delta^2) \quad (3.4)$$

When $k = 2$, the expression $\sum \gamma_j (1 - \gamma_j) (\pi_j + \pi_{j+1})$ reduces to the familiar formula $p(1 - p)$ for the binomial variance. There are many equivalent forms of this expression. The following are a few.

$$(i) \quad \sum_{j=1}^{k-1} \gamma_j (1 - \gamma_j) (\pi_j + \pi_{j+1})$$

$$(ii) \quad \sum_{j=1}^k \pi_j (1 - \gamma_j - \gamma_{j-1})^2$$

$$(iii) \quad \sum_{j=1}^{k-1} \gamma_j \gamma_{j+1} \pi_{j+1}$$

$$(iv) \quad \sum_{j=1}^{k-1} (1 - \gamma_j) (1 - \gamma_{j-1}) \pi_j$$

$$(v) \quad \frac{1}{3} - \frac{1}{3} \sum_{j=1}^k \pi_j^3$$

For continuous distributions, the analogue of these expressions is obtained by replacing γ by $F(X)$, π by $dF(X)$ and the summation becomes an integral. All are equal to $\frac{1}{3}$. Expression (v) shows that the variance of $\tilde{\Delta}$ is smallest when all categories have equal probability as we would expect.

A further slight problem is that the weights (3.3) are not obtainable directly and must be estimated from the data. Clayton (1974) used the weights

$$w_j = R_{.j} (n - R_{.j}) (n_{.j} + n_{.j+1})$$

which depend only on the column totals. Simulation results by McCullagh (1977) indicate that, for a wide range of conditions, these weights do not produce noticeable bias in $\tilde{\Delta}$. The variance estimator (3.4), with π_j estimated by $R_{.j}/n$, can however, be seriously misleading when $|\Delta| > 1$.

Clayton also considered a second estimator Δ^* which is consistent but appears to have a larger bias than $\tilde{\Delta}$. The second estimator is

$$\Delta^* = \log\left[\frac{\{\sum w_j^* R_{2j}(n_1 - R_{1j})\}}{\{\sum w_j^* R_{1j}(n_2 - R_{2j})\}}\right]$$

where w_j^* are weights estimated by $(n_{.j} + n_{.j+1})$. Both $\tilde{\Delta}$ and Δ^* have the same asymptotic variance (3.4) but Δ^* has an inconsistency which makes it difficult to generalise to the case of several groups. Thus, if there are three groups, the logistic differences $\Delta_{12}, \Delta_{13}, \Delta_{23}$ must satisfy the consistency property $\Delta_{13} = \Delta_{12} + \Delta_{23}$. This property of the parameters is satisfied by the estimator $\tilde{\Delta}$ but not by Δ^* . For this reason, Δ^* is not considered further. However, if there are only two groups, Δ^* may have some computational advantages over $\tilde{\Delta}$ especially for programmable calculators.

3.3 Maximum Likelihood Estimation.

When $k = 2$ the maximum likelihood estimator, $\hat{\Delta}$ is the log of the sample cross ratio. For $k > 2$ the maximum likelihood estimator of Δ cannot be obtained explicitly but can be obtained by iterative numerical procedures. For details of the general case see 54.4.

Some properties of $\tilde{\Delta}$ can be obtained from the matrix of second derivatives of the log likelihood function. McCullagh (1977) has shown that the asymptotic variance of $\hat{\Delta}$ is given by (3.4). Hence, for small Δ , and provided that the counts are sufficiently large, there is no loss of efficiency incurred by using $\tilde{\Delta}$ rather than $\hat{\Delta}$. In particular, tests based on $\tilde{\Delta}$ of the hypothesis that $\Delta = 0$ have asymptotic relative efficiency equal to one compared with tests based on $\hat{\Delta}$.

The simplicity and closed form of $\tilde{\Delta}$ make it preferable to $\hat{\Delta}$ for some purposes. However the maximum likelihood method provides estimates of all the parameters so that the adequacy of the model can be checked by a χ^2 test. Other possible checks include plots of standardised residuals and tests based on residuals.

3.4 An Example.

The following data from Holmes and Williams (1954) are a classification of 1398 children aged 0-15 years according to tonsil size and whether or not they were carriers of *Streptococcus pyogenes*. These data have been discussed and analysed by several authors: see for example Armitage (1955), Cox (1970) p. 11 and Clayton (1974).

Table 3

Tonsil sizes of carriers and non-carriers of *Streptococcus pyogenes*

	Tonsil Size			Total
	Present but not enlarged	Enlarged	Greatly Enlarged	
Carriers	19	29	24	72
Non-carriers	497	560	269	1326
Total	516	589	293	1398

The sample logits are -1.009, 0.683 for carriers and -.511, 1.367 for non-carriers. The weights are 0.638, 0.362 giving $\tilde{\Delta} = .565$ with standard error 0.225. The parameter Δ provides strong evidence that tonsil sizes tend to be larger in the carrier group than in the non-carrier group.

To check the adequacy of the linear logistic model, all parameters were estimated by maximum likelihood giving

$$\hat{\Delta} = .603 \pm .225$$

$$\hat{\theta}_1 = -.810 \pm .116$$

$$\hat{\theta}_2 = 1.061 \pm .118$$

The residual deviance or likelihood ratio chi-squared statistic is $G^2 = 0.302$ on one degree of freedom indicating a good fit.

An alternative method of analysing these data is to regard the proportion of carriers as the dependent variable and tonsil size as the explanatory variable. The usual procedure, since the explanatory variable is ordered is assign the scores $-1, 0, 1$ to the tonsil size categories. Since the dependent variable is binary the usual linear logistic model with the scores as a covariate can be fitted. See, however, the comments in §2.7.

An alternative to assigning arbitrary scores to the categories is to use the method of isotonic regression, Barlow, Bartholomew, Bremner and Brunk (1972). This method does not require scores to be assigned to the factor levels but it does take account of their order. Note that isotonic regression deals with ordered factors whereas the present paper is concerned with the case where the possible response categories are ordered.

A small point to notice is that $\tilde{\Delta}$ is generally smaller in absolute value than $\hat{\Delta}$. This shrinkage occurs because of the addition of $\frac{1}{2}$ to the numerator and denominator of (3.2). One possibility for reducing the bias of $\hat{\Delta}$ is to add $\frac{1}{2}$ to the end categories in each row. In this example, the addition of $\frac{1}{2}$ to categories 1 and 3 reduces $\hat{\Delta}$ to .600.

Another probability for reducing bias is to use the jackknifed estimate of Δ which eliminates terms of order n_1^{-1}, n_2^{-1} in the bias of $\hat{\Delta}$ and simultaneously gives an alternative estimate of variance. To apply the jackknife in this case we reduce each cell count in turn by unity and compute the maximum likelihood estimates for each of the six reduced tables. Let $\hat{\Delta}_{(ij)}$ be the estimate

of Δ when the $(1,j)$ cell is reduced by one. Then, under the assumption that the bias in $\hat{\Delta}$ is of order $a_1/n_1 + a_2/n_2$ where a_1 and a_2 do not depend on n , it is easily shown that the estimate

$$\Delta^J = (n_1 + n_2 - 1)\hat{\Delta} - \sum_{i=1}^2 (n_i - 1)/n_i \sum_{j=1}^k n_{ij} \hat{\Delta}_{(ij)} \quad (3.5)$$

has a lower order of bias than $\hat{\Delta}$. Similar comments apply to $\hat{\theta}$ but it is not clear whether the bias in $\hat{\theta}$ is of the same order as that in $\hat{\Delta}$. The extension of (3.5) to several samples is straightforward and eliminates bias of the order $\Sigma a_i/n_i$. Unfortunately, the bias in more complex examples is not generally of the form $\Sigma a_i/n_i$ so that the generalisation of (3.5) does not remove first order bias (Farewell, 1978). Hinkley (1977) gives a general discussion of jackknifing in unbalanced problems.

Generalisations of the usual variance estimator are not so straightforward, but we argue as follows. Let Y_{i1} , $i = 1, \dots, n_1$ be independent and identically distributed multinomial random variables. It is convenient to think of the Y 's as taking values $1, 2, \dots, k$ rather than being vector valued. Similarly, let Y_{i2} , $i = 1, \dots, n_2$, the corresponding observations in the second sample, be independent, identically distributed though not necessarily distributed similarly to those in the first sample. The samples are assumed to be mutually independent. Suppose $\hat{\Delta}$ were expressible as a difference of the form

$$\hat{\Delta} = \frac{1}{n_1} \sum_{i=1}^{n_1} g_1(Y_{i1}) - \frac{1}{n_2} \sum_{i=1}^{n_2} g_2(Y_{i2}) \quad (3.6)$$

where g_1, g_2 are arbitrary but sufficiently smooth functions. There is good reason — see §3.2 — for believing that $\hat{\Delta}$ is at least approximately of the form (3.6). Then under (3.6)

$$\text{var}(\hat{\Delta}) = \text{var}\{g_1(Y_1)\}/n_1 + \text{var}\{g_2(Y_2)\}/n_2. \quad (3.7)$$

given the independence assumptions.

Let $\hat{\Delta}_1(-1)$ be the estimate of Δ when the i^{th} observation is removed from the first sample and define the pseudo values for the first sample by

$$n_1 \hat{\Delta} - (n_1 - 1) \hat{\Delta}_1(-1) \quad i = 1, \dots, n_1.$$

Similarly, the pseudo values for the second sample are

$$n_2 \hat{\Delta} - (n_2 - 1) \hat{\Delta}_2(-1) \quad i = 1, \dots, n_2.$$

Under (3.6), the pseudo values in the first sample are equal to

$$g_1(Y_{11}) - \frac{1}{n_2} \sum g_2(Y_{12})$$

while those for the second sample are

$$- g_2(Y_{12}) + \frac{1}{n_1} \sum g_1(Y_{11}).$$

Thus, when $\hat{\Delta}$ is expressible in the linear form (3.6) an unbiased estimate of $\text{var}(\hat{\Delta})$ can be obtained by combining the 'within' sum of squares for the pseudo values in each group. This analysis suggests that even if $\hat{\Delta}$ is not quite linear the variance estimator

$$\sum_i (n_i - 1) / n_i \sum_j n_{ij} (\hat{\Delta}_{(ij)} - \bar{\hat{\Delta}}_{(i)})^2 \quad (3.8)$$

is approximately unbiased for $\text{var}(\hat{\Delta})$ and hence for $\text{var}(\Delta^J)$ since $\Delta^J = \hat{\Delta}$ under the linearity assumption (3.6). The above analysis extends readily to the many sample case.

Rounding errors can cause serious problems in using the jackknife, particularly if the cell counts are large. For example, the usual criterion of four correct significant digits is quite inadequate in the present example. This is because the difference $\hat{\Delta}_{(ij)} - \hat{\Delta}$ is very small when the (ij) cell count is large. One possibility is to increase the accuracy of the estimates but this requires extra fitting cycles with a corresponding increase in cost. Another possibility is to reduce cell counts by an amount proportional to their size and to use the continuity of the estimates to calculate an approximation to $\hat{\Delta}_{(ij)} - \hat{\Delta}$. The second possibility has not yet been fully explored.

Application of the above technique to the present example gave the following estimates and standard errors.

$$\Delta^J = .601 \pm .236$$

$$\theta_1^J = -.808 \pm .122$$

$$\theta_2^J = 1.060 \pm .123$$

In virtually all examples considered jackknifing had little effect on the estimates while the standard errors were usually about 5% larger than those obtained from the asymptotic covariance matrix. For those reasons plus the considerable expense of computing jackknifed estimates for a large table, jackknifing is not considered worthwhile in the present context.

4.1 The General Linear Logistic Model.

We now consider the problem of estimating the parameters of the general linear logistic model (2.4) which in its saturated form can be written

$$\lambda_{ij} = \theta_j - \alpha_i \quad \begin{matrix} j = 1, \dots, k-1 \\ i = 1, \dots, t \end{matrix} \quad (4.1)$$

and in unsaturated form

$$\lambda_{ij} = \theta_j - \beta^T \tilde{x}_i \quad (4.2)$$

where $\tilde{\beta} = (\beta_1, \dots, \beta_p)$ is a vector of unknown parameters and $p < t - 1$. Although the location part of model (4.1) is saturated since there are only $t - 1$ degrees of freedom among t location parameters, the model does not fit the data exactly except when $k = 2$. Consequently, for $k > 2$, the adequacy of the saturated location model can be tested in the usual way.

Two complementary approaches to estimation are taken. The first involves defining a generalised empirical logistic transform and using the transformed variable to estimate $\tilde{\beta}$ by weighted least squares. The advantages of this method are

- (i) the estimate of $\tilde{\beta}$ always exists and is approximately unbiased,
 - (ii) the estimate can be obtained non-iteratively
- and (iii) standard linear regression programs can be used.

The main disadvantage is that no estimate of the parameter $\tilde{\theta}$ is obtained and hence it is not possible to assess the adequacy of the logit assumption.

The second approach is via maximum likelihood and the advantages of the method are well known. The main disadvantage is that a special iterative computer algorithm is needed to obtain the estimates but the adequacy of the logit assumption can be checked by a chi-squared test or by an examination of the residuals.

4.2 A Generalised Empirical Logistic Transform.

In the two sample case the estimator $\tilde{\Delta}_{21}$ of the logistic difference between group 2 and group 1 was defined by $\tilde{\Delta}_{21} = \sum w_j \tilde{\lambda}_{2j} - \sum w_j \tilde{\lambda}_{1j}$ where $\sum w_j = 1$ and $w_j \propto R_j(n_j - R_j)(n_j + n_{j+1})$. The quantity

$$Z_i = \sum w_j \tilde{\lambda}_{ij} \quad (4.3)$$

with weights $\{w_j\}$ as given above is called the generalised empirical logistic transform of the observations in the i^{th} group. The estimator $\tilde{\Delta}_{21}$ is simply $Z_2 - Z_1$. By the arguments used in §3.2,

$$E(Z_i) = \sum w_j \theta_j - \beta^T X_i + O(n_i^{-2}) \quad (4.4)$$

and
$$\text{Var}(Z_i) = [n_i \sum \gamma_j (1 - \gamma_j) (\pi_j + \pi_{j+1})]^{-1} + O(|\beta|). \quad (4.5).$$

For fixed weights $\{w_j\}$, the transformed variables $\{Z_i\}$ are independent. As noted in §3.2, the randomness in the weights does not appear to induce noticeable correlations among the $\{Z_i\}$. Consequently, the transformed variables $\{Z_i\}$ can be regarded as independent and approximately normally distributed with mean given by (4.4) and variance given by (4.5).

If, however, the logistic difference between groups is large we may wish to improve on the variance estimator (4.5). We now derive a formula for $\text{var}(Z)$ which is approximately unbiased even when the logistic differences are large. Consider a typical row with cumulative sums $\{R_j\}$ and let $\{w_j\}$ be a set of arbitrary but fixed positive weights with $\sum w_j = 1$. Let the $k - 1$

sample logits be $\{\tilde{\lambda}_j\}$. Numerical evidence by Gart and Zweifel (1967) supports the estimation of $\text{var}(\tilde{\lambda}_j)$ by

$$\tilde{v}_{jj} = \frac{n+1}{(R_j + \frac{1}{2})(n - R_j + \frac{1}{2})} \quad (4.6)$$

This is such that, to a close approximation, $E(\tilde{v}_{jj}) = \text{var}(\tilde{\lambda}_j)$. The covariance of $\tilde{\lambda}_i$ and $\tilde{\lambda}_j$ for $i \leq j$ can be estimated by

$$\tilde{v}_{ij} = \frac{n+1}{(n - R_i + \frac{1}{2})(R_j + \frac{1}{2})} \quad i \leq j \quad (4.7)$$

Numerical evidence for this estimator is provided in Appendix 1.

Hence an approximately unbiased estimate of the covariance matrix of $\{\tilde{\lambda}_j\}$ is given by the Green's matrix $\tilde{V}$ with elements $\tilde{v}_{ij}$ given by (4.6) and (4.7). The variance of the transformed variable Z_i is estimated by

$$\text{Var}(Z_i) = \tilde{w}_i^T \tilde{V}_i \tilde{w}_i \quad (4.8)$$

The requirement that $\tilde{V}$ be nearly unbiased is particularly appropriate here since the final variance of interest is a weighted combination of variances and covariances of $\{\tilde{\lambda}_i\}$.

4.3 An Example.

We use the disturbed dreams data from Table 1 to illustrate the application of the generalised empirical logistic transform. The column totals are 40, 41, 42, 100 giving weights .180, .290, .530 and $\Sigma \gamma_j (1 - \gamma_j) (\pi_j + \pi_{j+1})$ is estimated as 0.297. According to formula (4.5) the variance of Z_i is simply $[0.297n_i]^{-1}$ where n_i is the total for the i^{th} age group. To compute the variance of Z_2 according to formula (4.8) we set up the covariance matrix

$$\tilde{V}_2 = 50 \begin{bmatrix} (13.5 \ 36.5)^{-1} \\ (24.5 \ 36.5)^{-1} (24.5 \ 25.5)^{-1} \\ (39.5 \ 36.5)^{-1} (39.5 \ 25.5)^{-1} (39.5 \ 10.5)^{-1} \end{bmatrix}$$

and find $\tilde{w}^T \tilde{V}_2 \tilde{w} = 0.0716$. This procedure is repeated for each age group yielding the values in Table 4.

Table 4

Empirical logistic transform and variance estimates

Age group	Z	Variance estimate	
		(4.5)	(4.8)
5 - 7	.2045	.1603	.1543
8 - 9	.5119	.0687	.0716
10 - 11	.3964	.0673	.0667
12 - 13	-.3741	.0571	.0571
14 - 15	-1.4181	.0765	.1158

Apart from age group 14 - 15 the discrepancy between the variance formulae is less than 5%. The discrepancy in the last age group could have been expected because the value for Z in this group is very much less than that for any other group.

A weighted linear regression of Z on negative age (using category mid points) with weights given by the reciprocals of (4.8) gives a regression coefficient of $\cdot 201$ with standard error 0.050 . The weighted residual sum of squares is 5.90 on 3 degrees of freedom, corresponding to a significance level of about 15%. Thus a linear regression on the logit scale appears adequate to explain the data although there is some evidence that the change with age may not be exactly linear on the logit scale.

Similar conclusions are reached if the weights are based on (4.5). In this case the estimate of β is 0.217 with standard error $\cdot 047$. The weighted residual sum of squares is 6.43 on 3 degrees of freedom.

It is instructive to compare the results of the preceding analysis with the maximum likelihood analysis in §4.5.

4.4.1 A Maximum Likelihood Analysis.

Models analogous to (2.4) and (2.5) but using the integrated normal response function are considered by Aitchison and Silvey (1957), Ashford (1959) and Finney (1971), p. 220 who give a maximum likelihood analysis. More general response functions are considered by Gurland, Lee and Dolan (1960) who give an analysis by weighted least squares. Although the general discussion throughout this paper refers to the logistic response function, the analysis in this section is more general as arbitrary response functions are considered.

The technical problem of fitting the model is considerably simplified by noting that the structure of the problem is that of a multivariate generalised linear model. Univariate generalised linear models have been discussed by Nelder and Wedderburn (1972). The contribution to the likelihood function of a single multinomial observation is $\pi_1^{n_1} \dots \pi_k^{n_k}$ where $\sum \pi_j = 1$ and the π_j 's depend on their covariates. Since the categories are ordered it is reasonable to take the cumulative transformation:

$$\begin{array}{lll} nZ_1 = R_1 = n_1 & \gamma_1 = \pi_1 \\ nZ_2 = R_2 = R_1 + n_2 & \gamma_2 = \gamma_1 + \pi_2 \\ \text{etc.} & \text{etc.} \\ Z_k = 1 & R_k = n & \gamma_k = \pi_k \end{array}$$

In terms of the cumulatives, the contribution to the likelihood can be written

$$\left[\left(\frac{\gamma_1}{\gamma_2} \right)^{R_1} \left(\frac{\gamma_2 - \gamma_1}{\gamma_2} \right)^{R_2 - R_1} \right] \left[\left(\frac{\gamma_2}{\gamma_3} \right)^{R_2} \left(\frac{\gamma_3 - \gamma_2}{\gamma_3} \right)^{R_3 - R_2} \right] \dots \left[\left(\frac{\gamma_{k-1}}{\gamma_k} \right)^{R_{k-1}} \left(\frac{\gamma_k - \gamma_{k-1}}{\gamma_k} \right)^{R_k - R_{k-1}} \right].$$

These factors are respectively, the probability given R_2 , that the first two cells divide in the ratio $R_1 : R_2 - R_1$; the probability given R_3 that the proportion in cell 3 relative to cells 1 and 2 combined is $R_3 - R_2 : R_2$ etc.

$$\text{Define } \phi_j = \log\{\gamma_j / \gamma_{j+1} - \gamma_j\} = \text{logit}(\gamma_j / \gamma_{j+1})$$

$$\text{and } g(\phi) = \log\{1 + \exp(\phi)\} = \log\{\gamma_{j+1} / (\gamma_{j+1} - \gamma_j)\}.$$

The log likelihood can be written

$$l = n[\{z_1 \phi_1 - z_2 g(\phi_1)\} + \{z_2 \phi_2 - z_3 g(\phi_2)\} + \dots + \{z_{k-1} \phi_{k-1} - g(\phi_{k-1})\}] \quad (4.9)$$

A univariate generalised linear model contains only one of the above components.

The following relationships follow from (4.9).

$$E(Z_j | Z_{j+1}) = Z_{j+1} g'(\phi_j) = Z_{j+1} \gamma_j / \gamma_{j+1}$$

$$E(Z_j) = \gamma_j$$

$$\text{Var}(Z_j | Z_{j+1}) = \frac{1}{n} Z_{j+1} g''(\phi_j)$$

$$\text{and } \text{Var}(Z_j) = \frac{1}{n} \gamma_j (1 - \gamma_j).$$

Analogous to the theory of univariate generalised linear models, let

$$v_j = \frac{\partial \gamma_j}{\partial \phi_j} = \gamma_{j+1} g''(\phi_j).$$

The conditional variance of Z_j given Z_{j+1} is $\frac{1}{n} v_j Z_{j+1} / \gamma_{j+1}$ so that $\frac{1}{n} v_j$ is the average value of the conditional variance of Z_j given Z_{j+1} .

4.4.2 The Link Function.

The link function — usually logit or probit for quantal regression — relates the unconditional means, γ_j to a structural model linear in the parameters.

$$Y_j = \text{Link}(\gamma_j) = \theta_j - \beta^T \underline{x}_j = \beta^{*T} \underline{x}_j^*$$

where Y_j is called the linear predictor for γ_j , $\beta^* = (\theta, \beta)$ and $\underline{x}_j^* = (0, \dots, 0, 1, 0, \dots, 0, X)$ where the 1 occurs in column j .

The derivatives of the log likelihood with respect to β_r^* are

$$\frac{\partial \ell}{\partial \beta_r^*} = \sum_{j=1}^{k-1} \frac{\partial \ell}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{dY_j} x_{jr}^* + \frac{\partial \phi_{j+1}}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} x_{j+1r}^* \right\}$$

where $\frac{\partial \ell}{\partial \phi_j} = n\{Z_j - Z_{j+1}\gamma_j/\gamma_{j+1}\}$.

Since $\frac{\partial \phi_j}{\partial \gamma_j} = v_j^{-1}$ and $\frac{\partial \phi_{j+1}}{\partial \gamma_{j+1}} = -v_j^{-1}\gamma_j/\gamma_{j+1}$ it follows that

$$\frac{\partial \ell}{\partial \beta_r^*} = \sum_{j=1}^{k-1} \frac{\partial \ell}{\partial \phi_j} v_j^{-1} q_{jr} \quad (4.10)$$

where $q_{jr} = \left\{ \frac{\partial \gamma_j}{\partial Y_j} x_{jr}^* - \gamma_j/\gamma_{j+1} \frac{d\gamma_{j+1}}{dY_{j+1}} x_{j+1r}^* \right\}$.

The second derivatives can be shown to be

$$-E \left(\frac{\partial^2 \ell}{\partial \beta_r^* \partial \beta_s^*} \right) = n \sum_{j=1}^{k-1} v_j^{-1} q_{jr} q_{js} \quad (4.11)$$

Expanding $\frac{\partial \ell}{\partial \beta^*}$ by a Taylor series about $\beta^* = \hat{\beta}^*$ and using expected second derivatives, we have

$$\frac{\partial \ell}{\partial \beta^*} = \frac{\partial \ell}{\partial \beta^*} \bigg|_{\beta^* = \hat{\beta}^*} + A \delta \beta^* + \dots$$

where the elements of A are given by (4.11). Hence the Newton-Raphson method with expected second derivatives is equivalent to iterating on the equation.

$$A \beta_{n+1}^* = A \beta_n^* + \frac{\partial \ell}{\partial \beta} = b \quad (4.12)$$

where the elements of b are given by

$$(b)_r = n[V_j^{-1} q_{jr} \left[y_j \frac{dy_j}{dy_j} + z_j - y_j/y_{j+1} \right] y_{j+1} \frac{dy_{j+1}}{dy_{j+1}} + z_{j+1} \bigg] \quad (4.13)$$

Expressions (4.11) and (4.13) are the contributions to A and b of a single multinomial sample. Hence, to obtain the estimate of β^* we obtain the total contribution to A and b and iterate on (4.12).

Convergence is quadratic and therefore extremely fast. Initial estimates of θ can be obtained from the column totals, but to keep these estimates stable and centred around zero it is better to regress on $X_i - \bar{X}$ rather than on X_i . Initial estimates of β can be obtained by the method of §4.2 but convergence is usually rapid even when these are set to zero. The only serious problem which could occur arises from the restriction that the $\{\theta_j\}$ be monotone increasing. A very bad initial estimate could lead to inversions of θ 's and hence to negative fitted values. However, with sensible initial estimates this problem rarely arises.

4.5 An Example.

We again use the data of table 1 to illustrate the maximum likelihood method. Although there is a choice of response functions we use the logistic to keep the analysis comparable to that in §4.3.

The convergence criterion ensures that all parameter estimates are accurate to at least four significant digits. Initial estimates of β are set to zero while θ is estimated from the column totals under the assumption that $\beta = 0$. Convergence occurred after four cycles, the last being used to verify the accuracy of the previous cycle. The maximum likelihood estimates of the parameters in the structural model $\lambda_{ij} = \theta_j - \beta(t_i - \bar{t})$ are

$$\hat{\theta}_1 = -1.498 \pm .178$$

$$\hat{\theta}_2 = -.493 \pm .144$$

$$\hat{\theta}_3 = .331 \pm .142$$

$$\hat{\beta} = .219 \pm .0495$$

where, as before, the age-category mid points are used.

For this model, the likelihood ratio chi-squared statistic is $G_1^2 = 12.42$ on 11 degrees of freedom. To assess the adequacy of the linear regression on age, the location model was saturated by including a row factor with 5 levels. The saturated location model is $\lambda_{ij} = \theta_j - \alpha_i$ and the deviance is reduced to $G_2^2 = 7.15$ on 8 degrees of freedom. Thus the test statistic for deviations from linearity, $G_1^2 - G_2^2 = 5.27$ on 3 degrees of freedom is very close to the weighted residual sum of squares in §4.3 and the conclusions from the maximum likelihood

analysis are similar to those from the least squares analysis. Finally, to test the adequacy of the logit assumption we use G^2_2 which is approximately equal to its expected value under the model. Hence the linear model on the logit scale seems adequate to explain the data. Examination of residuals reveals no unexpected features.

5.1 The Non-linear Logistic Model.

The linear logistic model has an interpretation in terms of an underlying continuum with contiguous intervals representing the categories. An important assumption leading to the linear model is that the variance or scale parameter on the logit scale remain constant in each row of the table. Ultimately this is an empirical assumption and, where possible, ought to be checked. One method of verification is to fit a more general model allowing the scale parameter to vary in a systematic way from row to row. The unequal scale parameters, if they are needed, should not be regarded merely as a ploy to obtain a better fit to the data. The scale parameters may contain substantive interpretable information in terms of degree of agreement or consensus: such information is not available from an examination of the location part of the model alone.

The multiplicative logistic model allowing arbitrary scale parameters $\{\tau_i\}$ associated with each row is

$$\lambda_{ij} = \log\{y_{ij}/(1 - y_{ij})\} = (\theta_j - \beta_j^T x_i)/\tau_i. \quad (5.1)$$

Since there is one scale parameter associated with each multinomial observation, model (5.1) is saturated in scale parameters. It is, however, unsaturated in location parameters provided $\dim(\beta_j) < t - 1$. To make the scale parameters identifiable it is convenient to impose a constraint such as $\tau_1 = 1$ or $\sum \ln \tau_i = 0$. The latter convention is adopted here and is particularly appropriate when we consider unsaturated scale models satisfying

$$\log \tau_i = \tau^T (x_i - \bar{x}), \quad (5.2)$$

where $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)^T$ and $\underline{\tau} = (\tau_1, \dots, \tau_p)^T$ is a vector of unknown parameters. Finally, there is no requirement that the covariates in (5.2) coincide with the covariates in the location part of (5.1).

5.2 Fitting the Non-linear Logistic Model

One method of obtaining parameter estimates in the non-linear logistic model is to use the generalised empirical logistic transform (4.3) with estimated variance given by (4.8) and estimate the parameters $(\underline{\beta}, \underline{\tau})$ by non-linear weighted least squares. This method has several disadvantages. Firstly, if the location model is saturated the transformed variables are fitted exactly: consequently the transformed variables contain no information on the remaining parameters of the model. Secondly, even when the total number of parameters in the location and scale models is less than $t - 1$, i.e. $\dim(\underline{\beta}, \underline{\tau}) < t - 1$, there is likely to be appreciable loss of information about the scale parameters in particular. This loss is caused by considering only one linear combination of the λ_{ij} and ignoring the other possible combinations. The combination considered in (4.3) is constructed to preserve location information and some scale information is inevitably lost. Thirdly, it is generally the case that the non-linear model is used because the linear model gives an unsatisfactory fit. We will, therefore, wish to check the adequacy of various aspects of the non-linear model. These checks are most easily carried out when fitted values and test statistics are available.

Note that the above comments do not apply to binary data ($k = 2$) since the empirical logit transform, together with its variance estimate contain all the information in the data. When $k > 2$ some information is lost although, as was shown in §3, the information loss in the linear model is quite small.

An alternative and asymptotically fully efficient method of fitting is based on the sample logits $\tilde{\lambda}_{ij}$ and their block-diagonal approximate covariance matrix with elements given by the matrices $\{\tilde{V}_t\}$ of §4.2. While this method is computationally feasible it will not be attempted here.

We now describe two methods for obtaining maximum likelihood estimates of the parameters in the nonlinear model (5.1) where the scale values satisfy (5.2). The first method involves fixing the scale parameters τ and noting that for fixed τ the location parameters (θ, β) can be estimated by the method of §4.4 with weights n_i/τ_i instead of n_i . Secondly, for fixed location parameters, the scale parameters τ can be estimated by the same method but using a logit link and an offset of $\log(\theta_j - \beta^T \tilde{x}_j)$ since

$$\log \lambda_{ij} = \log(\theta_j - \beta^T \tilde{x}_j) - \tau^T (\tilde{x}_j - \bar{\tilde{x}}) .$$

The offset can be imaginary but an algorithm can be designed so that complex numbers do not occur. The system iterates back and forth until convergence occurs. The method is not recommended since often a large number of cycles are required for convergence and, in addition, the method yields approximate conditional covariance matrices of the parameters only. These conditional variances and covariances are not appropriate for the usual tests of significance.

The second method involves computing the complete matrix of second derivatives and using the Newton-Raphson process as described in §4.4. Details of the method for arbitrary link functions are given in Appendix 2. For the second method, convergence is generally rapid and the maximum likelihood estimates are obtained when they exist and are finite. For details of existence and uniqueness see §6.1.

5.3 An Example.

We use the data of table 2 to illustrate an application of the non-linear model. The model of no association yields a deviance or likelihood ratio chi-squared statistic of $G_0^2 = 85.2$ on 16 degrees of freedom. Saturating the location model reduces the value of G^2 to $G_1^2 = 49.1$ on 12 degrees of freedom. Clearly, the linear model does not fit the data. Patterns of large residuals for treatments 4 and 5 indicate that an unequal scales model may be appropriate. Saturating the scale model reduces the deviance to $G_2^2 = 21.3$ on 8 degrees of freedom.

Although the formal test of comparing the difference $G_1^2 - G_2^2 = 27.8$ on 4 degrees of freedom is not strictly valid since the larger model does not fit the data, it is clear that the unequal scales model is a substantial improvement over the linear model. The following summary table is often a useful aid to interpretation.

Table 5

Logistic summary statistics for the taste-testing data

<u>Treatment</u>	<u>Location</u>	<u>Scale</u>	<u>Deviance contribution</u>
1	.07	1.25	1.86
2	.56	1.05	5.30
3	-1.11	.90	1.97
4	-.50	1.66	11.62
5	.98	.51	0.59
			<u>21.34</u>

It is clear that the major contributor to the total deviance is treatment 4. Examination of residuals reveals that, for treatment 4, many judges reacted extremely, giving either very low or very high scores. This diversity of opinion is also reflected in the large scale value for this observation. The anomalous reaction of the judges to treatment 4 was also noted by Snell (1964) who used a similar model but did not allow for unequal variances.

If treatment 4 were set aside, a satisfactory fit would be obtained for the remainder of the data. However, the conclusions for the reduced data set are essentially the same as those based on the summary statistics in table 5. To summarise, treatment 5 has the most favourable overall response and the consensus among judges is also highest for this treatment ($\tau_5 = .50$). Treatment 3 rates worst overall and the consensus for this is also high ($\tau_3 = .90$). As we might have expected, agreement among the judges is greatest when the treatment is either very good or very bad. This is not an artefact of the observations piling up in the end categories.

Further detailed analysis including multiple comparisons among the location and scale parameters is possible with the aid of the asymptotic covariance matrix. Standardised residuals can be checked for unexpected features but these follow-up procedures are not pursued here.

6.1 Existence and Uniqueness of Maximum Likelihood Estimates.

In this section we show that under certain limiting conditions, unique and finite maximum likelihood estimates exist. The first of these conditions is that the model be identifiable. Let the covariates in the location model be $X_1, \dots, X_r$ and those in the scale part be $U_1, \dots, U_s$. Identifiability requires that (a) $r < t$, $s < t$ and $r + s < (k - 1)t$ and (b) the matrices $\underline{X} = [\underline{X}_1, \dots, \underline{X}_r, \underline{1}]$ and $\underline{U} = [\underline{U}_1, \dots, \underline{U}_s, \underline{1}]$ have full rank $r + 1$ and $s + 1$ respectively where $\underline{1}$ is the unit vector of length t . The further requirement that the scale values $\{\tau_i\}$ be positive for each row is implied by the linear parametrisation in terms of $\log \tau_i$.

Theorem 4.1 of Berk (1972) can be adapted to show that, provided that the row totals $\{n_i\}$ tend to infinity in fixed proportion, then the probability tends to one that a unique maximum of the likelihood function exists. This conclusion depends on certain monotonicity properties of the link function. For the particular model studied here, the preceding result follows in a relatively straightforward way. We first note from the formulae of Appendix 2 that the negative matrix of second derivatives, $\underline{H}$, of the log likelihood function has positive semi-definite expectation $\underline{A}$. This is true for all admissible parameter values since $\underline{A} = \sum \underline{A}_i$ with $\underline{A}_i = \underline{Q}_i \underline{Q}_i^T$ where $\underline{Q}_i$ is a full rank matrix of order $p \times k - 1$. Since the random part of $\underline{H}$ is of smaller order than $\underline{A}$, it follows that $\underline{H}$ is almost surely positive semi-definite. Furthermore, if the model is identifiable, $\underline{A}$ is strictly positive definite with eigenvalues bounded away from zero as the number of observations becomes large. Hence the likelihood surface is almost surely strictly concave and, since there can be no asymptotes, a unique maximum must exist.

While this result is reassuring, its practical importance is fairly limited. We therefore investigate some properties of the Fisher scoring algorithm of Appendix 2. The central equation of this algorithm can be written.

$$\underline{A}_{n-n+1} \underline{\psi}_{n+1} = \underline{A}_{n-n} \underline{\psi}_n + \left. \frac{\partial \ell}{\partial \underline{\psi}} \right|_{\underline{\psi} = \underline{\psi}_n} \quad (6.1)$$

where $\underline{\psi}_1, \dots, \underline{\psi}_n$ is the sequence of parameter estimates and $\underline{A}_n$ is the expected matrix of second derivatives at $\underline{\psi}_n$. Since $\underline{A}$ has full rank it follows that

$$\underline{\psi}_{n+1} = \underline{\psi}_n \quad \Rightarrow \quad \left. \frac{\partial \ell}{\partial \underline{\psi}} \right|_{\underline{\psi} = \underline{\psi}_n} = 0$$

and

$$\left. \frac{\partial \ell}{\partial \underline{\psi}} \right|_{\underline{\psi} = \underline{\psi}_n} = 0 \Rightarrow \underline{\psi}_{n+1} = \underline{\psi}_n$$

Hence all stationary points of the likelihood function are fixed points of (6.1) and vice-versa. However, it seems possible that there could be several fixed points of (6.1) corresponding to multiple stationary points of the likelihood function. As far as I can tell no such pathological examples are known. We now consider a few special cases where the maximum likelihood estimates are infinite.

Suppose that the saturated linear model is being fitted and two of the rows in the table have one of the following patterns with X representing a non-zero observation

Row A: X X X 0 0 or X 0 0 0
 Row B: 0 0 0 X X X X X X .

Since the logit (or probit etc.) difference between rows A and B is free to vary, the estimate of this difference will be infinite.

A second problem occurs when the saturated scale model is being used. If all the observations in a particular row fall into the same category, the estimate of the scale parameter for that row will be 0 ($-\infty$ on the log scale) which is on the boundary of the parameter space. For either of these problems to arise it is necessary only that the relevant parameter be free to vary. This occurs when the model — location or scale — is saturated but it can occur even when the appropriate model is unsaturated.

A third problem which can occur when the data are sparse is that a particular category contains no observations. Formally, this means that, for some j , $\hat{\theta}_j = \hat{\theta}_{j+1}$. Numerical difficulties may arise however but these are easily remedied by deleting the appropriate category. A computer package developed at the University of Chicago for fitting the models described in this paper automatically deletes unused categories and has options for merging adjacent categories with small counts.

Experience with this package indicates that while the estimates may be infinite under a variety of conditions in addition to those cases discussed above, there have been no uniqueness problems even in small samples when finite estimates exist. It is conjectured, therefore, that, under conditions similar to those described by Haberman (1977), the maximum likelihood estimates are unique when they exist and are finite.

References

- Abelson, R. P. and Tukey, J. W. (1963) Efficient utilisation of non-numerical information in qualitative analysis: general theory and the case of simple order. *Ann. Math. Statist.*, 34, 1347-1369.
- Aitchison, J. and Silvey, S. D. (1957) The generalisation of probit analysis to the case of multiple responses. *Biometrika*, 44, 131-140.
- Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375-386.
- Ashford, J. R. (1959) An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics*, 15, 573-581.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1973) *Statistical Inference Under Order Restrictions*. New York, Wiley.
- Bartlett, M. S. (1935) Contingency table interactions. *J. Roy. Statist. Soc. Suppl.* 2, 248-252.
- Berk, R. H. (1972) Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.*, 43, 193-204.
- Bradley, R. A., Katti, S. K. and Coons, I. J. (1962) Optimal scaling for ordered categories. *Psychometrika*, 27, 355-374.
- Clayton, D. G. (1974) Some odds-ratio statistics for the analysis of ordered categorical data. *Biometrika*, 61, 525-531.
- Cox, D. R. (1970) *The Analysis of Binary Data*. London, Chapman and Hall.

- Cox, D. R. (1972) Regression models and life tables. (With discussion).
J. Roy. Statist. Assoc. (B), 34, 187-220.
- Farewell, V. T. (1978) Jackknife estimation with structural data. Biometrika,
65, 444-447.
- Fienberg, S. E. (1975) The observational study — a review. J. Amer. Statist.
Assoc., 70, 521-523.
- Finney, D. J. (1971) Probit Analysis. 3rd edition, Cambridge University Press.
- Fisher, R. A. (1963) Statistical Methods for Research Workers. 13th edition.
Edinburgh, Oliver and Boyd.
- Gart, J. J. and Zweifel, J. R. (1967) On the bias of various estimators of the
logit and its variance with application to quantal bioassay. Biometrika, 54,
181-187.
- Goodman, L. A. (1964) Interaction in multidimensional contingency tables.
Ann. Math. Statist., 35, 632-646.
- Goodman, L. A. (1978) Simple models for the analysis of association in cross-
classifications having ordered row and ordered column categories. Technical
Report No. 61. Dept. of Statistics, University of Chicago.
- Goodman, L. A. and Kruskal, W. H. (1954) Measures of association for cross-
classifications. J. Amer. Statist. Assoc., 49, 732-764.
- Gurland, J., Lee, I., and Dolan, P. A. (1960) Polychotomous quantal response in
biological assay. Biometrics, 16, 382-398.

Haberman, S. J. (1974) Log-linear models for frequency tables with ordered classifications. *Biometrics*, 30, 589-600.

Haberman, S. J. (1977) Maximum likelihood estimates in exponential response models. *Ann. Statist.*, 5, 815-841.

Haitovsky, Y. (1973) Regression Estimation from Grouped Observations. London, Griffin.

Healy, M. J. and Goldstein, H. (1976) An approach to the scaling of categorised attributes. *Biometrika*, 63, 219-229.

Hinkley, D. V. (1977) Jackknifing in unbalanced situations. *Technometrics*, 19, 285-292.

Holmes, M. C. and Williams, R. E. O. (1954) The distribution of carriers of *Streptococcus pyogenes* among 2413 healthy children. *J. Hyg. Camb.*, 52, 165-179.

Mandel, J. (1959) The analysis of Latin squares with a certain type of row-column interaction. *Technometrics*, 1, 379-387.

Mandel, J. (1971) A new analysis of variance model for non-additive data. *Technometrics* 13, 1-18.

Maxwell, A. E. (1961) Analysing Qualitative Data. London, Methuen.

McCullagh, P. (1977) Analysis of Ordered Categorical Data. Ph.D. Thesis. Univ. of London.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalised linear models. *J. Roy. Statist. Assoc. (A)*, 135, 370-384.

Nishisato, S. and Arri, P. S. Nonlinear programming approach to optimal scaling of partially ordered categories. *Psychometrika*, 40, 525-548.

Pearson, K. (1913) Note on the surface of constant association. *Biometrika* 9, 534-537.

Plackett, R. L. (1965) A class of bivariate distributions. *J. Amer. Statist. Assoc.*, 60, 516-522.

Plackett, R. L. (1974) *The Analysis of Categorical Data*. London, Griffin.

Savage, I. R. (1957) Nonparametric Statistics. *J. Amer. Statist. Assoc.*, 52, 331-344.

Simon, G. (1974) Alternative Analyses for the singly ordered contingency table. *J. Amer. Statist. Assoc.*, 69, 971-976.

Snell, E. J. (1964) A scaling procedure for ordered categorical data. *Biometrics*, 20, 592-607.

Stevens, S. S. (1951) Mathematics, measurement and psychophysics. In *Handbook of Experimental Psychology*. S. S. Stevens Editor. New York: Wiley.

Stevens, S. S. (1958) Problems and methods of psychophysics. *Psychological Bulletin* 55, 177-196

Stevens, S. S. (1968) Measurement, Statistics and the Schemapiric view. *Science*, 161, 849-856.

Tukey, J. W. (1954) Unsolved problems of experimental statistics. J. Amer. Statist. Assoc., 49, 706-731.

Williams and Grizzle (1972) Analysis of contingency tables having ordered response categories. J. Amer. Statist. Assoc., 67, 55-63.

Yates, F. (1948) The analysis of contingency tables with groupings based on quantitative characters. Biometrika 35, 176-181.

Yule, G. U. (1912) On the methods of measuring association between two attributes. J. Roy. Statist. Assoc., 75, 579-652.

APPENDIX 1

ON THE COVARIANCE OF SAMPLE LOGITS

Let $n_1, \dots, n_k$ be a multinomial sample with cumulative totals $R_1, \dots, R_k = n$. The j^{th} sample logit is defined as $\tilde{\lambda}_j = \ln\{(R_j + \frac{1}{2}) / (n - R_j + \frac{1}{2})\}$, $j = 1, \dots, k - 1$. We now consider numerically two asymptotically equivalent estimators $v_{ij}^{(1)}$ and $v_{ij}^{(2)}$ of $\text{cov}(\tilde{\lambda}_i, \tilde{\lambda}_j)$. These are

$$v_{ij}^{(1)} = \frac{n+1}{(R_j + \frac{1}{2})(n - R_i + \frac{1}{2})} \quad 1 \leq j < k,$$

which is a natural generalisation of Gart's variance estimator, and

$$v_{ij}^{(2)} = \frac{n+1}{n} \left[\frac{n+2}{(R_j + 1)(n - R_i + 1)} \right] \quad 1 \leq j < k,$$

which is a similar generalisation of the variance estimator due to Goodman (1964). Goodman, however, was instrumental in the development of both variance estimators.

For numerical purposes it is sufficient to consider $k = 3$ and the values of n considered are $n = 5, 10, 20, 40$. A wide range of multinomial probability vectors π could be chosen. Symmetry considerations reduce the choice a little. The following multinomial vectors were chosen to represent a variety of possibilities.

$$\pi_1 = (0.1, 0.2, 0.7)$$

$$\pi_2 = (0.1, 0.6, 0.3)$$

$$\pi_3 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$$

$$\pi_4 = (0.2, 0.4, 0.4)$$

$$\pi_5 = (0.25, 0.05, 0.70)$$

The following table gives the exact percent bias of $\tilde{v}^{(1)}$ $\tilde{v}^{(2)}$ for four values of n and five values of π .

Table 6

Exact percent bias of two estimators of the covariance of sample logits.

n	$\tilde{v}^{(1)}$			$\tilde{v}^{(2)}$		
	$v_{11}^{(1)}$	$v_{22}^{(1)}$	$v_{12}^{(1)}$	$v_{11}^{(2)}$	$v_{22}^{(2)}$	$v_{12}^{(2)}$
$\pi_1 = (0.1, 0.2, 0.7)$						
5	187.2	23.6	152.8	103.6	5.9	115.1
10	84.8	3.9	55.8	27.2	-5.5	41.0
20	26.7	-1.1	12.3	-5.8	-3.3	9.0
40	3.3	-1.1	1.1	-11.3	-1.2	0.2
$\pi_2 = (0.1, 0.6, 0.3)$						
5	187.2	23.6	131.0	103.6	5.9	153.7
10	84.8	3.9	43.1	27.2	-5.5	51.9
20	26.7	-1.1	10.4	-5.8	-3.3	14.1
40	3.3	-1.1	1.2	-11.3	-1.2	3.0
$\pi_3 = (\frac{1}{2}, \frac{1}{3}, \frac{1}{3})$						
5	16.9	16.9	20.5	2.7	2.7	25.7
10	2.0	2.0	3.5	-4.8	-4.8	7.3
20	-2	-2	1.4	-2.3	-2.3	3.7
40	-0	-0	.8	-8	-8	2.0
$\pi_4 = (0.2, 0.4, 0.4)$						
5	61.8	8.5	42.8	27.0	-9	49.2
10	19.4	0.2	9.9	-3.2	-3.2	14.2
20	1.9	-1	1.8	-8.2	-1.0	4.2
40	-1	-0	.9	-3.8	-3	2.1
$\pi_5 = (0.25, 0.05, 0.70)$						
5	38.2	23.6	35.6	13.5	5.9	16.1
10	9.2	3.9	7.7	-5.6	-5.5	-2.1
20	0.2	-1	0.4	-5.5	-3.3	-2.7
40	-1	-1	0.2	-2.2	-1.2	-9

Table 6 supports the following general conclusions:

- (i) $\underline{v}^{(1)}$ is almost always conservative (biased upwards). Negative biases, where they occur, are extremely small;
- (ii) The negative biases of $\underline{v}^{(2)}$ can be substantial even when the mean value per cell is 4 or more;
- (iii) For small expected cell counts $\underline{v}^{(2)}$ is less biased than $\underline{v}^{(1)}$ but in such cases neither estimator could be considered satisfactory;
- (iv) Otherwise $\underline{v}^{(1)}$ is generally preferable to $\underline{v}^{(2)}$;
- (v) For $0.1 < \gamma_1 < \gamma_2 < 0.9$ and $\min\{n\gamma_1, n(1 - \gamma_2)\} > 4$ the bias of $\underline{v}^{(1)}$ is less than 3.3%;
- (vi) Small expected cell counts in the middle category have little effect on the bias of $\underline{v}^{(1)}$ provided that the end categories have expected cell counts greater than about 4.

APPENDIX 2

FITTING THE GENERAL NON-LINEAR MODEL

The general nonlinear model can be written

$$y_j = \text{link}(\gamma_j) = \beta^{*T} X_j^* \exp(\tau^T U)$$

where β^* , X_j^* are as defined in §4.4.2 and U_i , the vector of covariates in the scale model, is normalised so that $\Sigma U_i = 0$. Let $\psi^T = (\beta^{*T}, \tau^T)$ be the vector of unknown parameters and $w = \exp(\tau^T U)$.

The derivative of the log likelihood with respect to β^* is

$$\frac{\partial \ell}{\partial \beta_r^*} = w \sum_{j=1}^{k-1} \frac{\partial \ell}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{dY_j} X_{jr}^* + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} X_{j+1,r}^* \right\}.$$

Substituting $V_j = \frac{\partial \gamma_j}{\partial \phi_j}$ and $\frac{\partial \phi_j}{\partial \gamma_{j+1}} = \frac{-\gamma_j}{\gamma_{j+1}} V_j^{-1}$ we obtain

$$\frac{\partial \ell}{\partial \beta_r^*} = w \sum_{j=1}^{k-1} \frac{\partial \ell}{\partial \phi_j} V_j^{-1} q_{jr}$$

where $q_{jr} = \left\{ \frac{d\gamma_j}{dY_j} X_{jr}^* - \frac{\gamma_j}{\gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} X_{j+1,r}^* \right\}$. Similarly the expected second derivative is

$$A_{rs} = -E \left(\frac{\partial^2 \ell}{\partial \beta_r^* \partial \beta_s^*} \right) = n w^2 \sum_j V_j^{-1} q_{jr} q_{js}$$

For the scale parameter τ the derivatives are

$$\frac{\partial \ell}{\partial \tau_r} = \sum \frac{\partial \ell}{\partial \phi_j} \left\{ \frac{\partial \phi_j}{\partial \gamma_j} \frac{d\gamma_j}{dY_j} Y_j U_r + \frac{\partial \phi_j}{\partial \gamma_{j+1}} \frac{d\gamma_{j+1}}{dY_{j+1}} Y_{j+1} U_r \right\}$$

which reduce to

$$\frac{\partial \ell}{\partial \tau_r} = U_r \sum \frac{\partial \ell}{\partial \phi_j} V_j^{-1} q_j$$

$$\text{where } q_j = Y_j \frac{d\gamma_j}{dY_j} - \frac{\gamma_j}{Y_{j+1}} Y_{j+1} \frac{d\gamma_{j+1}}{dY_{j+1}}.$$

Similarly the expected second derivatives are

$$-E \left(\frac{\partial^2 \ell}{\partial \tau_r \partial \tau_s} \right) = n U_r U_s \sum V_j^{-1} q_j^2,$$

and the mixed second derivatives are

$$-E \left(\frac{\partial^2 \ell}{\partial \beta_r^* \partial \tau_s} \right) = n w U_s \sum V_j^{-1} q_j q_{jr},$$

The negative matrix of expected second derivatives is therefore positive semi-definite for all admissible parameter values.

The Taylor series expansion for $\frac{\partial \ell}{\partial \psi}$ gives

$$\frac{\partial \ell}{\partial \psi} = \frac{\partial \ell}{\partial \psi} \bigg|_{\psi = \hat{\psi}} + A(\delta \psi).$$

where $\delta\psi = \hat{\psi} - \psi$ and A is the expected matrix of second derivatives. Hence, updated values of ψ are obtained by iterating on the equation

$$\psi_{n+1} = \psi_n + \frac{\partial \ell}{\partial \psi} = b$$

Writing $W = \log w = \tau^T U$ we have

$$\begin{aligned} (A\psi)_r &= nw \sum_j V_j^{-1} q_{jr} \sum_s q_{js} \beta_s + nw \sum_s U_s \tau_s \sum_j V_j^{-1} q_{jr} q_j \\ &= nw(1+W) \sum_j V_j^{-1} q_{jr} q_j \end{aligned}$$

for $r \leq \dim(\beta^*)$. Similarly, for $s > \dim(\beta^*)$,

$$\begin{aligned} (A\phi)_s &= nWU_s \sum_j V_j^{-1} q_j^2 + nU_s \sum_j V_j^{-1} q_j^2 \\ &= n(1+W)U_s \sum_j V_j^{-1} q_j^2. \end{aligned}$$

The corresponding elements of b are given by

$$(b)_r = nw \sum_j V_j^{-1} q_{jr} q_j (1+W) + z_j - \frac{\gamma_j}{\gamma_{j+1}} z_{j+1}$$

and

$$(b)_s = nU_s \sum_j V_j^{-1} q_j q_j (1+W) + z_j - \frac{\gamma_j}{\gamma_{j+1}} z_{j+1}.$$

Finally, all the above equations represent the contribution to the log likelihood and its derivatives of a single multinomial observation and the total contribution is the sum of the individual contributions.